WISE: Weighted Iterative Society-of-Experts for Multimodal Multi-Agent Debate with Probabilistic Consensus

Anoop Cherian [1] **Suhas Lohit** [1] **Kuan-Chuan Peng** [1]

Abstract

Multi-agent debate (MAD) is a powerful paradigm for combining multiple large language models (LLMs) to achieve robust reasoning, but prior work has largely focused on language-only settings, leaving its multimodal potential underexplored. We present Weighted Iterative Society-of-Experts (WISE), a generalized MAD framework that systematically integrates heterogeneous multimodal LLMs to address challenging vision-and-language tasks in a zero-shot setting. Our key idea is to factor agents into three roles based on their multimodal capabilities: *Solvers*, which process multimodal inputs and generate candidate solutions; *Reflectors*, which may or may not access multimodal inputs but evaluate solutions, provide feedback, and assign weights; and an *Orchestrator*, which operates unimodally to reason over solutions and feedback and produce directives that guide subsequent reasoning. To account for varying agent reliability, we introduce an unsupervised probabilistic aggregation method, termed WISE–Dawid–Skene, which leverages the weighting scheme in WISE-MAD to adaptively combine agent outputs. We evaluate WISE on several challenging mathematical reasoning datasets and show that it consistently outperforms state-of-the-art methods across diverse LLM configurations, demonstrating its effectiveness as a general and scalable multimodal reasoning framework.

1. Introduction

In recent years, large language models (LLMs) have been widely adopted as agents for a broad spectrum of tasks, spanning everyday applications to domains requiring advanced capabilities such as scientific discovery and mathematical reasoning (Luo et al., 2025; Taylor et al., 2022; Ahn et al., 2024). A central challenge in making LLM agents effective for reasoning tasks lies in ensuring that their chains of thought (CoT) are both logical and correct. However, due to a range of internal and external factors—including spurious data correlations, sensitivity to sampling temperature, computational errors, and context-length limitations—their predictions are often brittle.

To address this, a variety of ensembling mechanisms have been explored (Chen et al., 2023; Wang et al., 2022; Shinn et al., 2023; Liang et al., 2023). Multi-agent debate (MAD) (Du et al., 2023; Liang et al., 2023) generalizes these approaches by allowing multiple LLMs to critique and refine each other's solutions, thereby reducing errors and improving final responses. While MAD has shown considerable benefits in purely language-based tasks, where LLM agents are relatively mature, its potential for improving multimodal LLMs (MLLMs) remains largely unexplored. Given that MLLMs still lag significantly behind LLMs on popular benchmarks, we investigate how MAD can improve multimodal problem solving, specifically those involving vision and language.

As multimodal LLMs (MLLMs) remain less reliable on visual and other non-linguistic inputs than on text, a central challenge is to recover correct solutions despite systematically unreliable and heterogeneous agent responses. Prior language-only approaches address this through repeated sampling and feedback, such as self-consistency (Wang et al., 2022) and reflective refinement (Shinn et al., 2023; Feng et al., 2024; Chen et al., 2023; Valmeekam et al., 2023). While these methods can be extended to multimodal settings, they largely treat all candidate solutions as exchangeable and do not account for modality-dependent error patterns (e.g., misperception, grounding failures, or inconsistent cross-modal reasoning). As a result, they lack mechanisms to identify and downweight unreliable agents or failure modes that are specific to multimodal reasoning. For instance, Figure 1 shows a simple multimodal problem (Cherian et al., 2024) and attempts by state-of-the-art MLLMs. Solving such problems requires task-specific knowledge, reflection from multiple perspectives, and targeted cues for reconsideration until the solution is well grounded.

Building on this insight and the classical "society of agents"

[1]Mitsubishi Electric Research Labs (MERL), Cambridge, MA. Correspondence to: Anoop Cherian <cherian@merl.com>.

Published as a workshop paper at SCALE - ICML 2026

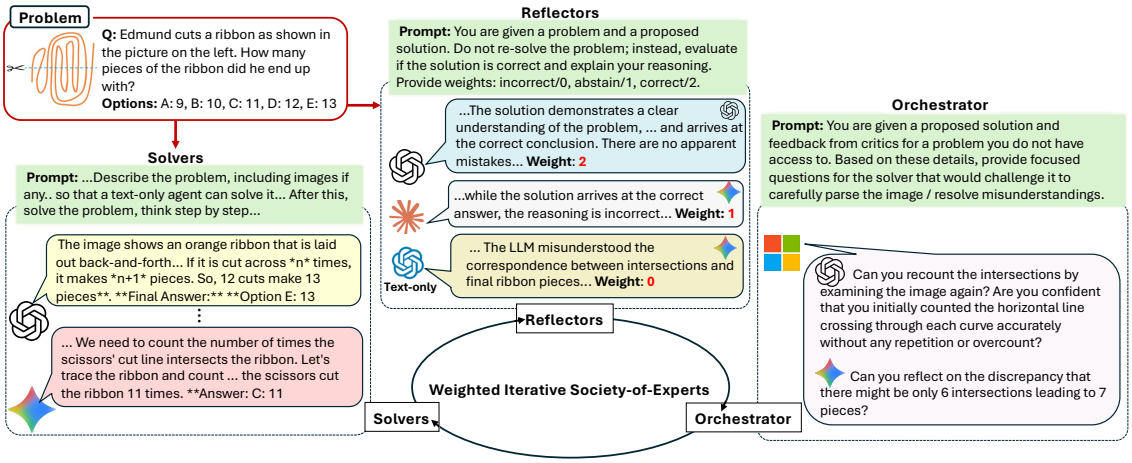


Figure 1. An illustration of our Weighted Iterative Society-of-Experts architecture for heterogeneous multi-agent debate where the agents are factorized into three distinct roles: solvers, reflectors, and orchestrators. We show the control flow and the prompts used for each role.

view from cognitive science, we introduce Weighted Iterative Society-of-Experts (WISE), a generalized multimodal MAD framework that enables heterogeneous agents to collaborate on complex problems. Each agent is pretrained on diverse corpora for specific downstream tasks (e.g., mathematical reasoning, coding, general knowledge), providing complementary capabilities that can be jointly leveraged.

A natural approach to designing WISE is to adopt frameworks such as RECONCILE (Chen et al., 2023), where agents are placed in a debate and report their confidence in their proposed solutions. While intuitive, this strategy faces several limitations in the multimodal setting: (i) it assumes all agents have similar multimodal capabilities, (ii) modern MLLMs are often heavily trained and tend to be overly confident—even when hallucinating, (iii) agents with diverse abilities may excel at solving tasks but not at critiquing (or vice versa), and (iv) involving all agents in debate can generate a quadratic number of messages, causing significant computational overhead.

To address these challenges, WISE organizes agents into three complementary roles: Solvers, Reflectors, and an Orchestrator. Solvers generate candidate reasoning chains over multimodal inputs, while Reflectors evaluate these solutions by verifying correctness, assigning discrete weights, and providing structured feedback. The Orchestrator coordinates this process by aggregating reflector feedback, summarizing critiques, and posing targeted queries that guide subsequent rounds of reasoning (see Figure 1). This iterative interaction not only improves error correction and robustness, but also naturally produces a collection of weighted, interdependent candidate solutions across agents and rounds. This structure motivates a principled approach to solution aggregation that explicitly models uncertainty and agent reliability. To this end, we build on the Dawid–Skene model (Dawid & Skene, 1979), treating agents as noisy annotators with latent, role-dependent error patterns. We extend this framework to the multi-agent debate setting by allowing errors to evolve across rounds and by coupling solver-generated responses with reflector-assigned weights. This leads to WISE-DS, a probabilistic model that performs expectation maximization over a mixture of multinomial distributions, jointly inferring agent-specific error matrices and posterior solution probabilities across both solver and reflector agents. Overall, WISE and WISE-DS together provide a unified, reliability-aware reasoning framework, where the multi-agent design and probabilistic aggregation are tightly integrated rather than treated as separate components.

To evaluate the efficacy of WISE, we conduct experiments on three datasets: (i) SMART-840 (Cherian et al., 2024), (ii) VisualPuzzles (Song et al., 2025), and (iii) EvoChart-QA (Huang et al., 2025). The first two focus on vision-and-language (VL) reasoning in a multiple-choice format, while the latter involves free-form short answers. Across these datasets and varied MAD configurations, we show that WISE delivers consistent improvements, beating the state-of-the-art results with substantial margins (2–7%).

We summarize our key contributions below:

1. **Setup:** We study MAD for zero-shot multimodal reasoning problems, a novel and unexplored setting. We propose a multimodal MAD framework that partitions agents into possibly overlapping *Solvers* and *Reflectors*, a previously unexplored aspect.
2. **Solution Aggregation:** We propose the WISE-Dawid–Skene algorithm for estimating agent error probabilities and deriving consensus solutions.
3. **Results:** WISE outperforms the state-of-the-art performance across three VL benchmarks, covering both multiple-choice and free-form answer formats.

2. Related Works

Self-Correction Methods. These methods improve LLM responses either by sampling multiple outputs (Wang et al., 2022; Liang et al., 2023) or by incorporating self/external feedback. Sampling-based methods seek robustness through majority voting, while feedback-based methods assume that reflection enables models to identify and fix their own errors. Representative methods include Self-Refine (Madaan et al., 2023), Self-Correction (Welleck et al., 2022), Reflexion (Shinn et al., 2023), Recursively Criticize and Improve (Kim et al., 2023), RL from AI Feedback (Bai et al., 2022), and others (Ganguli et al., 2023; Jiang et al., 2023; Dhuliawala et al., 2023; Pan et al., 2023). However, empirical studies show that without external feedback, self-correction often fails to improve—and can even harm—performance (Huang et al., 2023; Kamoi et al., 2024; Valmeekam et al., 2023).

Multi-Agent Debate. MAD generalizes correction and feedback by enabling agents to debate solutions, e.g., (Du et al., 2023) studies debates among identical LLMs, while (Liang et al., 2023) introduces external feedback and a judge model to manage the process. Heterogeneous setups include collaboration/competition schemes (Feng et al., 2024), secretary-style orchestration (Wang et al., 2024b), and RECONCILE (Chen et al., 2023), where agents share uncertainty scores and iteratively refine answers. Explain-Analyze-Generate (Gu et al., 2025) highlights risks of misleading agents and advocates task decomposition. (Subramaniam et al., 2025) fine-tunes MAD models for improved convergence. In contrast, WISE introduces explicit agent roles, avoids reliance on self-estimated confidence (which often needs external calibration (Thind et al., 2025; Khan et al., 2024)), and naturally supports abstention through weighted scoring.

Mixture of LLMs. Ensemble-based methods combine multiple LLMs in various ways. Mixture-of-Agents (MoA) (Wang et al.) treats ensembles as layers controlled by prompts and gating models, while Self-MoA (Li et al., 2025a) selects only top-performing outputs for downstream processing. Other works explore heterogeneous mixtures resolved by majority voting (Li et al., 2024), or orchestrated systems such as Magnetic-One (Fourney et al., 2024). Unlike them, WISE embeds agents in an iterative self-reflective loop, using feedback to progressively refine solutions. There have also been approaches that trains mixtures-of-experts (Li et al., 2025b; Wu et al., 2024; Shen et al., 2024), however usually involve model training.

Multimodal Methods. MAD has rarely been studied in the multimodal setting. For example, (Wang et al., 2025a) uses multimodal MAD for knowledge transfer to smaller models but does not address vision–language reasoning tasks. Process reward models have been applied for multimodal reasoning (Wang et al., 2025b), and multi-domain reward models have also been explored (Zeng et al., 2025). To our knowledge, WISE is the first to systematically study MAD for multimodal reasoning with heterogeneous agents.

Aggregation Approaches. Most MAD systems aggregate solutions via majority voting (Du et al., 2023) or weighted averaging using confidence estimates (Chen et al., 2023). From a crowdsourcing perspective, agents can be viewed as workers prone to systematic errors, motivating the use of quality-control methods for statistically sound consensus estimation (Dawid & Skene, 1979; Hovy et al., 2013; Ma & Olshevsky, 2020; Majdi & Rodriguez, 2023; Ustalov et al., 2021). WISE extends this line by adapting Dawid–Skene to account for both Solver and Reflector roles, while incorporating confidence weights.

3. Proposed Method – WISE

3.1. Problem Setup

Given a problem p from a multimodal domain $\mathcal{P}$, we aim to design a model $M : \mathcal{P} \to \mathcal{A}$ that outputs the correct answer $a \in \mathcal{A}$. We focus on vision–language problems, where each p is an image–question pair requiring visuo-linguistic understanding. We denote the model's prediction by $\hat{a} = M(p)$ and seek $\hat{a} = a$. For multiple-choice problems, a and $\hat{a}$ are among K options; for free-form answers, we use an external model to assess equality.

A standard machine learning way to approach our problem setup is to consider a dataset with pairs of problems and their ground truth solutions, where M is trained to minimize the prediction error. We assume neither training data nor in-context examples are available, instead have access to a set $\mathcal{L}$ of heterogeneous LLMs (either unimodal or multimodal) that have been pretrained on diverse corpora of language or multimodal tasks, not necessarily the tasks in $\mathcal{P}$. Our goal is to use $\mathcal{L}$ to design M to solve $\mathcal{P}$ in a zero-shot setting.

Our key insight is that different LLMs/MLLMs, trained on different data, have complementary strengths: for instance, Qwen-VL (Wang et al., 2024a) targets general vision-and-language tasks, Phi-4 (Abdin et al., 2024) focuses on coding and math reasoning, and GPT-4.1[1] aims to bridge vision, language, and coding. Because it is difficult to know in advance which skills a problem requires—and thus which single LLM configuration is best—we instead let models with diverse capabilities debate different aspects of the problem, reach a consensus, and then apply a reliable aggregation scheme to obtain a plausible correct solution. Guided by this intuition, we propose our Weighted Iterative Society-of-Experts (WISE) architecture, illustrated in Fig. 2.

[1] https://openai.com/index/gpt-4-1/

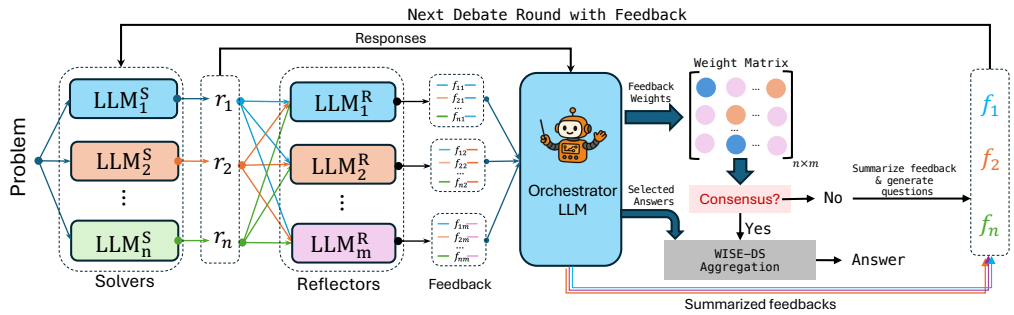


Figure 2. WISE architecture demonstrating the control flow, agent settings, feedback, responses, and the WISE-Dawid-Skene solution aggregation scheme to produce the final response.

3.2. WISE Architecture

Given a set of $|\mathcal{L}|$ unique agents, a standard MAD implementation (Du et al., 2023; Chen et al., 2023) needs agents to communicate with one another, resulting in nearly $|\mathcal{L}|^2$ messages. This quadratic growth can be redundant, costly when models are accessed through API calls, and computationally intensive when hosted locally. Moreover, the volume of messages may quickly exceed the input context size of the models. In the multimodal setting we consider, further restricting all agents to possess identical multimodal capabilities can be limiting; for example, some models may excel at language-based reasoning while others may have stronger vision–language alignment.

To address these challenges, WISE partitions the agents into two groups: a set of n Solvers $\mathcal{S} = \{\mathrm{L}_1^S, \mathrm{L}_2^S, \ldots, \mathrm{L}_n^S\}$ and a set of m Reflectors $\mathcal{R} = \{\mathrm{L}_1^R, \mathrm{L}_2^R, \ldots, \mathrm{L}_m^R\}$, where $\mathcal{S} \cup \mathcal{R} = \mathcal{L}$ and $\mathcal{S} \cap \mathcal{R}$ may be non-empty. The Solvers are assumed to be multimodal agents, while the Reflectors may include both unimodal and multimodal models.

Solver Models (First Round): For a problem p, each solver agent $\mathrm{L}_i^S \in \mathcal{S}$ receives as input the tuple $t_1^S = (\mathrm{PROMPT}_1^S, p, \mathcal{A}_p, f = \phi)$, where PROMPT_1^S is a textual prompt describing the task, and $\mathcal{A}_p$ denotes the candidate answer set (with $\mathcal{A}_p = \phi$ for free-form answers). The tuple also includes a feedback message f, which is null in the first round and populated with feedback from the reflectors and orchestrator in subsequent rounds. For vision–and–language problems, p also contains the associated images. The exact prompts used are provided in the Appendix, while essential properties of the prompts we use are depicted in Figure 1. Each solver L_i^S generates a response r_k^i to p at round k ($k = 1$ for the first round)[2]. The response is expected to include a description of the image (allowing non-visual models to understand the solution) and detailed explanations of the solution steps, along with the final selected answer from the candidate set, i.e., $\hat{a} \in \mathcal{A}_p$.

[2]We will drop k when the round index is not important.

Reflector Models: At the k-th round, a reflector $\mathrm{L}_j^R \in \mathcal{R}$ takes as input the tuple $t_k^R = (\mathrm{PROMPT}^R, p, r_k)$, where PROMPT^R specifies the reflector's task (see Figure 1). The task is twofold: (i) judge the correctness of the solver's response r_k, and (ii) provide textual feedback. For correctness, the reflector outputs a numerical weight $w_{ij}^k \in \{0, 1, 2\}$, with 0 for an incorrect answer, 1 for uncertainty/abstention, and 2 for a correct response. For feedback, the reflector produces a detailed explanation justifying its assigned weight, highlighting correct aspects and pointing out errors in r_k. Each solver's response is thus evaluated by all reflectors, yielding: (i) an $n \times m$ feedback matrix F^k whose (i, j)-th entry f_{ij}^k stores the textual feedback, and (ii) an $n \times m$ weight matrix W^k whose (i, j)-th entry is w_{ij}^k.

Orchestrator: Similar to (Chen et al., 2023; Wang et al., 2024b; Feng et al., 2024), a central component of WISE is the orchestrator agent O, implemented using an LLM, which governs the debate process and message exchange among agents. However, in contrast, WISE orchestrator has two extra responsibilities: (i) aggregating the feedback matrix F^k and weight matrix W^k to decide whether the debate should continue, and (ii) summarizing the collected feedback into actionable guidance for the solvers.

If the debate is to continue, the orchestrator produces consolidated feedback for each solver. Specifically, for solver i, it generates a feedback summary f_i^k as:

$$f_i^k = \mathrm{L}^O \left(\mathrm{PROMPT}^O, \bigoplus_{j \in [n], w_{ij} \neq -1} F_{ij}^k \right), i \in [n], \quad (1)$$

where L^O denotes the orchestrator's internal LLM, and the input consists of all non-null feedback from reflectors on solver i's response r^i, concatenated with the orchestrator prompt PROMPT^O. The prompt PROMPT^O instructs L^O to summarize the feedback and formulate a set of *challenge questions* that will force the solver to (re-)consider aspects of the image or language part of the problem in more detail, toward refining its solution in the next round (and thus we do not provide it access to the problem p).

The orchestrator then prepares a solver-specific feedback tuple $\left(\text{PROMPT}_k^S, p, \mathcal{A}_p, r^i \oplus f_i^k\right)$, where PROMPT_k^S is the updated solver prompt and $\oplus$ denotes concatenation of the solver's response with the orchestrator feedback. This tuple is forwarded to the i-th solver, and the process is repeated for all $i \in [n]$, thereby completing one debate round. The debate continues until either a maximum number of rounds is reached or consensus is achieved.

3.3. WISE Dawid-Skene Response Aggregation

As previously discussed, when employing heterogeneous models, a central challenge is to rigorously characterize their reliability. For instance, a model may generate effectively random outputs due to a high sampling temperature, may lack the requisite reasoning capabilities, or may exhibit other systematic deficiencies. Determining the extent to which such responses can be trusted is thus a nontrivial problem.

To address this issue, we draw upon the crowdsourcing literature (Majdi & Rodriguez, 2023), where a classical approach to calibrating annotator reliability is the Dawid–Skene (DS) model (Dawid & Skene, 1979). Given N workers, each providing one of K categorical labels for a given item, the DS model constitutes an unsupervised probabilistic framework that leverages expectation–maximization (EM) to jointly infer worker-specific error rates and the latent ground-truth label distribution (interpreted via a majority-consensus prior).

In our setting, we treat each (agent, round) pair as a distinct annotator endowed with its own error profile. This new formulation enables us to capture agents' ability to integrate critiques over rounds, their evolving error patterns, and the relative reliability of later-round responses. Accordingly, we instantiate our MAD formulation within the DS framework to model agent-specific error rates and to compute the posterior distribution over candidate solutions. This procedure attenuates the influence of agents that behave in a near-random manner and concomitantly amplifies the contribution of agents whose outputs exhibit consistent agreement with the emergent consensus.

Mathematically, suppose $p_{\alpha\beta}$ is the probability of selecting answer β for the true answer α by a model, and if ζ_α is the true prior probability of selecting an option α, then given Λ problems and the agents' responses, (Dawid & Skene, 1979) models the likelihood of the data as a mixture of multinomial distributions:

$$p(\text{data}) \propto \prod_{i=1}^{\Lambda} \sum_{\alpha=1}^{K} \zeta_\alpha \left[\prod_{j=1}^{N} \prod_{\beta=1}^{K} (p_{\alpha\beta}^j)^{\lambda_{i\beta}^j} \right], \tag{2}$$

where $\lambda_{i\beta}^j$ is the number of times j-th agent produced β as the answer against the true answer of α if the model is run multiple times on the same problem i. As both the true priors ζ_α and the error rates $p_{\alpha\beta}$ are unknown, DS uses EM to optimize the likelihood iteratively towards convergence.

In WISE, we have two sources of errors: i) solvers making errors in selecting the correct answers from the K choices and the reflectors selecting one of $J(=3)$ weights. Suppose $p_{\alpha_1\beta_1}^t$ and $p_{\alpha_2\beta_2}^c$ denote the error matrices for the solver and the reflectors respectively, in selecting option $\beta_1 \to \alpha_1$ (for $\alpha_1, \beta_1 \in [K]$) and selecting wrong weights for a provided answer, i.e., , selecting weight $\beta_2 \to \alpha_2$ (for $\alpha_2, \beta_2 \in [J]$), then the joint probability for estimating the combined error matrices and their joint true priors $\zeta_{\alpha\beta} = \zeta_\alpha \zeta_\beta$ could be modeled for debate round k as a product of two mixtures of multinomial distributions, given by:

$$\prod_{i=1}^{\Lambda} \sum_{\alpha=1}^{J} \sum_{\beta=1}^{K} \zeta_{\alpha\beta} \left[\prod_{c=1}^{|\mathcal{R}|} \prod_{\beta_2=1}^{J} \left(p_{\alpha_2\beta_2}^c\right)^{\lambda_{i\beta_2}^{R_c}} \prod_{t=1}^{|\mathcal{S}|} \prod_{\beta_1=1}^{K} \left(p_{\alpha_1\beta_1}^t\right)^{\lambda_{i\beta_1}^{S_t}} \right],$$

subject to $\sum_\alpha \zeta_\alpha = \sum_\beta \zeta_\beta = 1$. We use EM to estimate the joint error matrices and true probabilities, following which the posteriors are computed, which is then used to recompute the selected answers and the respective weights. See the Appendix for detailed derivations.

If $\widetilde{W}$ is the updated weight matrix given by the WISE-DS method above after the EM convergence, then we aggregate these weights across rounds towards finding the final answer. Specifically, given updated weight matrices from k rounds, we generate aggregated weight $\overline{W}$ for an answer $\hat{a}$ given by:

$$\overline{W}_{ij}^k(\hat{a}) = \sum_{\ell=1}^{k} \widetilde{W}_{ij}^\ell \frac{\ell}{k(k+1)}, \text{ if } r_i^\ell = \hat{a}. \tag{3}$$

The aggregated weighting scheme in equation 3 favors higher weights for more recent rounds while also normalizing them to [0,1] (recall that the weights in $\widetilde{W}$ are in $\{0, 1, 2\}$). We also maintain a distribution of weights across the answer candidates. Specifically, assume $|\mathcal{A}_p|$ updated answer options (after the DS step) for a given problem p, then the accumulated weight $\overline{w}_{\hat{a}}$ for option $\hat{a} \in \mathcal{A}_p$ as:

$$\overline{w}_{\hat{a}}^k = \sum_{i \in [n], j \in [m]} \overline{W}_{ij}^k(\hat{a}), \tag{4}$$

and we select the highest ranking answer a^* as the output from the final $\overline{w}$, where $a^* = \arg\max_{\hat{a}} \{\overline{w}_{\hat{a}}\}$. We apply the DS on a subset of the data towards estimating the error probabilities, following which we use them in the posterior estimates of the solutions on the full dataset.

4. Experiments

Datasets. We evaluate on three vision–language reasoning datasets: (i) SMART-840 (Cherian et al., 2024), (ii) Visual Puzzles (Song et al., 2025), and (iii) EvoChart-QA (Huang et al., 2025). SMART-840 has 840 children's

math Olympiad problems (grades 1–12), grouped into six consecutive grade-pair categories, supporting analysis by age level. Visual Puzzles has 1168 problems targeting algorithmic, analogical, deductive, inductive, and spatial multimodal reasoning. Both use multiple choice (5 options for SMART-840, 4 for Visual Puzzles). EvoChart-QA is the "complex" subset (320 problems) of ChartQA, with questions over line, bar, scatter, and pie charts; it uses free-form short answers and is considered the hardest, where SOTA MLLMs perform poorly.

Agents and Debate Notation. We use diverse LLM/MLLM agents in our experiments, listed below with their abbreviations: GPT-4o (G4o), GPT-4.1 (G41), GPT-5.2 (G52), o1-mini (o1m) o4-mini (o4m), Claude-Sonnet-3.5 (CS35), Claude-Sonnet-3.7 (CS37), Claude-Sonent-4.6 (CS46), Qwen2-VL-2.5-7B (QVL), Gemma3-8B (Ge3), and Phi4-3.4B ($\phi 4$). To represent a debate, we use the notation $(\mathrm{L}_1^S + \mathrm{L}_2^S + \cdots) \times (\mathrm{L}_1^R + \mathrm{L}_2^R + \cdots)|$orchestrator, where the first set is the solver models and the second set is the reflectors. For example, (G41+CS37)×(o4m+$\phi 4$)|G4o means we use GPT-4.1 and Claude-Sonnet-3.7 as the solvers, (o4-mini, Phi-4) as reflectors, and GPT-4o as the orchestrator. We use the compact notation $(\mathrm{L}_1 + \mathrm{L}_2 + \cdots)^2|O$ for a setup using $\mathrm{L}_1 + \mathrm{L}_2 + \cdots$ as both solvers and reflectors.

Debate Configuration: We ran up to 4 debate rounds, with early stopping when the orchestrator terminated the debate upon consensus. Our analysis considers debate setups inspired by prior work, using models closest to those previously studied. Wherever possible, we report performance from the original papers and supplement them with our own reproduced results. We stop WISE debate when all the reflectors agree the solution is correct or when the maximuim number of rounds is reached.

Evaluation and Implementation Details: We use multiple-choice accuracy on SMART-840 and VisualPuzzles, and use the Phi-4 model to compare predicted answers with ground truth on EvoChart-QA. WISE is implemented in PyTorch with MLLMs sourced from HuggingFace or accessed via API calls. Our experiments were conducted on an A100 node, assigning each MLLM to a dedicated GPU. The distributed setup employed asynchronous communication for message passing among MLLMs. We accessed GPT and Claude models via their APIs. Across all debate steps and models, we used the same prompts.

4.1. Experimental Results

SOTA Comparisons: In Tables 1, 2, and Fig. 3, we report results across the four datasets, comparing WISE against the SOTA methods. Overall, WISE consistently outperforms all baselines. On SMART-840, it improves accuracy by nearly 6% over RECONCILE (Chen et al., 2023), while also surpassing single-model, homogeneous, and self-correction variants, including the state-of-the-art G52 and CS46 models. On VisualPuzzles, WISE achieves a 2.8% gain with an ensemble of o4-mini, Claude-Sonnet-3.7, and Gemma3, demonstrating that benefits persist even with strong models such as o4-mini. On EvoChart-QA, WISE substantially outperforms individual models—improving GPT-4o (39.1%), Gemma3 (29%), and Qwen-VL (19.6%) to 52.6%. Using stronger models like Claude-Sonnet-3.7 and o4-mini further boosts performance to 75.4%, nearly 20% above previously reported results. These findings also highlight WISE 's effectiveness on free-form short-answer problems.

Method	LLMs / MLLMs	Image+Text	Text	Overall
Single	GPT-4.1	48.2	88.2	60.6
	Claude-Sonnet	35.3	71.0	44.3
	Gemma3	24.0	73.7	38.6
	GPT-5.2	65.0	95.9	72.0
	CS46	55.2	92.5	65.8
Self-Reflect (Shinn et al., 2023)	GPT-4.1	50.9	88.9	61.2
	Gemma3	30.0	58.2	46.0
Self-Consistency (Wang et al., 2022)	GPT-4.1	48.6	90.3	63.2
	Sonnet-3.5	35.3	68.4	44.2
	Gemma3	24.8	56.8	43.9
Homogeneous (Liang et al., 2023)	GPT-4.1	49.5	88.4	59.7
	Sonnet-3.5	34.9	71.0	45.2
	Gemma3	30.9	74.2	44.4
Single	G41+G4o+GS35+Ge3+QVL	45.5	82.7	55.6
Self-Reflect	Ge3+QVL+G41	42.6	83.4	53.9
Self-Consistency	G41+CS35+Ge3+QVL	47.2	84.1	56.6
Homogeneous	G41+CS35+Ge3+QVL	43.5	81.6	53.9
RECONCILE	G41 + Ge3 + S35	50.9	90.1	62.0
RECONCILE	G41+CS46+o4-m	62.7	97.3	73.1
WISE (ours)	$(\text{G41+Ge3+S35})^2$|G4o	**55.8**	**91.4**	**68.1**
WISE (ours)	$(\text{G52+CS46})^2$|G52	**70.6**	**98.2**	**77.4**

Table 1. Performance of MAD approaches on the SMART-840 dataset using varied configurations under similar number budget.

Method	LLMs/MLLMs	Alg.	Ana.	Ded.	Ind.	Spa.	Avg.*	Our Run†
Random	NA	25.0	25.0	25.0	25.0	25.0	25.0	NA
Human	NA	88.0	66.0	80.0	50.0	90.0	75.0	NA
Single	GPT-4o	49.2	58.3	49.0	27.3	26.2	41.3	42.1
	o4-mini	65.3	68.7	**75.5**	33.0	45.5	57.0	53.3
	CS37	64.5	48.3	65.0	26.8	37.4	48.3	46.6
	Gemma3	40.4	36.4	34.7	27.5	15.3	NA	31.6
WISE (ours)	$(\text{o4m+CS37+Ge3})^2$|G4o	**70.9**	**69.0**	75.1	**37.0**	**47.1**	**59.8**	**59.8**

Table 2. Comparisons on the Visual Puzzles Dataset. *indicates the overall performance reported in the original paper, unless it is our method. † indicates results obtained when we ran the model.

Table 3 compares our WISE-DS solution aggregation method with alternatives. Our DS extension yields consistent gains across debate configurations and datasets. For competing methods, we applied their aggregation scheme followed by weighted majority without the joint EM step. Overall, our approach surpasses the original Dawid–Skene model, while weighted majority voting further shows steady and promising improvements. In Appendix D, we show the error matrices and their correlation to agent's strength.

Dataset	LLMs/MLLMs	Maj. V	DS	Wt. Maj.V	WISE
SMART-840	$(\text{G41+CS35+Ge3})^2$|G4o	63.9	65.3	66.2	**68.1**
	(G41+S35)×(G4o+S35+$\phi 4$)|G4o	61.1	62.6	59.1	62.3
	(G4o+Q)×(Q+$\phi 4$+Ll3+Ge3)|$\phi 4$	49.1	49.2	50.0	**51.1**
	G4o×G4o|$\phi 4$	46.8	46.9	47.4	48.2
	$(\text{G52+CS46})^2$|G52	75.4	72.7	75.6	**77.4**
Visual Puzzles	$(\text{o4m+Ge3+C37})^2$|G4o	57.3	56.1	58.1	**59.8**
EvoChart-QA	$(\text{o4m+C37})^2$|Go	74.4	NA	**75.4**	NA

Table 3. Comparisons of WISE-DS vs. other methods. DS: Dawid-Skene, Maj. V: majority voting, Wt. Maj.V uses Eq. 3.

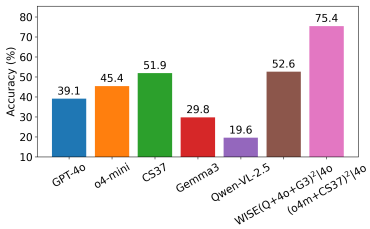


Figure 3. Accuracy (%) on the EvoChart-QA dataset.

Solvers	Reflectors	LLMs / MLLMs	Overall
Strong	Strong	(G41+S35)^2|G4o	67.7
Strong	Weak	(G41+S35)×(G4o+CH3+ϕ4)|G4o	61.7
Mix	Mix	(G41+Ge3+S35)^2|G4o	**68.1**
Weak	Weak	(G4o+QVL)×(Q+ϕ4+L3+G3)|G4o	49.7
Weak	L-only	G4o × (ϕ4+o1-mini)|G4o	51.5

Table 4. Performance of WISE using weak and strong models.

4.2. Ablation Studies

Multimodality? While our debate protocol is not inherently modality-specific, in Table 5, we explore if WISE's gains depend on multimodal information—and our results indicate that they do. To substantiate this, we conducted ablations by (i) removing images, (ii) scrambling (i.e., feeding unrelated images), and (iii) replacing images with captions generated by GPT-4.1 (G41) and Claude-Sonnet-4.6 (CS46). These experiments are performed on SMART-840 (grades 5–6) using GPT-5.2 (G52) and CS46, with standalone baselines reported. These results demonstrate that WISE's improvements are not modality-agnostic: they depend critically on access to and effective integration of multimodal cues.

Compute Budget? In Table 6, we report the full token-cost-performance analysis. The results are presented on the above settings. Relative to RECONCILE, WISE increases per-round cost only slightly while improving accuracy substantially. Likewise, our strongest WISE configuration is far above homogeneous, showing that the extra cost is converted into stronger reasoning than redundant model calls.

Agent's Reasoning Strength? In Table 4, we evaluate WISE across solver–reflector pairings: strong–strong, strong–mix, mix–mix, weak–weak, and weak–language-only, where model strength is defined by singleton performance (see Appendix). All experiments use GPT-4o as orchestrator and WISE-DS aggregation on SMART-840. Strong–strong performs well, but mix–mix slightly outperforms it (68.1 vs. 67.7), indicating that controlled heterogeneity is beneficial. Weak–weak performs worst, yet still improves over the best individual model (e.g., GPT-4o rises from 42.4% to 49.7% within WISE). Unlike RECONCILE, which assumes homogeneous capabilities, WISE supports heterogeneous role assignments and is advantageous in weak–language-only settings (51.5% accuracy, 1.1% above RECONCILE under comparable compute). While

Ablation	G52	CS46	WISE
Baseline	51.4	55.9	65.3
No images	41.6	39.1	42.6
Scrambling	43.8	41.8	47.9
G41 captions	50.7	55.8	55.8
CS46 captions	53.8	54.2	59.4

Table 5. Analyzes of multimodal properties of WISE debate using (G52+CS46)^2|G41 on a SMART-840 subset.

Config.	Debate	Tok./LLM	#Tok.	Cost	Latency	Acc. (%)
(G52+CS46)^2|G41	WISE	7215.35	1946.77	7.6	12.23	65.3
(G41+Ge3+S35)^2|G4o	WISE	6392.45	1617.94	4.8	10.95	58.2
(G41+Ge3+S35)	RECONCILE	5758.60	1447.93	4.4	5.93	50.0
G41+CS35+Ge3+Q3	Homogeneous	3050.69	771.28	2.6	12.96	41.3

Table 6. Comparison of configurations. We report the tokens/round, latency (s) / round and cost/round.

strong–strong configurations help WISE, heterogeneous pairings reveal that (i) strong–mix with weak reflectors underperforms due to poor feedback, and (ii) weak–weak and weak–language-only remain competitive, showing WISE's robustness to agent variability.

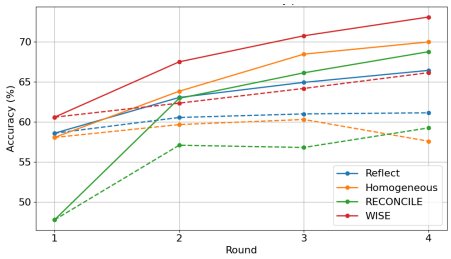


Figure 4. Plot of the cumulative and average accuracy for varied configurations against the number of debate rounds.

Accuracy across rounds? Fig. 4 plots cumulative accuracy (fraction of problems solved in round k) and average accuracy/round for multiple configurations. WISE performs strongly in both metrics. The close alignment of cumulative and average accuracy shows that our aggregation scheme is robust. In particular, the gap between cumulative accuracy (an upper bound) and average accuracy is much smaller for WISE-DS compared to other schemes.

5. Conclusions

We proposed Weighted Iterative Society-of-Experts (WISE), a novel formulation of multi-agent debate (MAD) involving heterogeneous LLMs and MLLMs with complementary capabilities. WISE supports multi-round multimodal debates by partitioning models into Solvers and Reflectors, coordinated by an orchestrator that manages the debate. Our results highlight several key findings: (i) MAD is effective in multimodal settings with appropriate debate designs, and (ii) segregating agents into roles enables effective use of abilities, leading to state-of-the-art results.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., and Yin, W. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*, 2024.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Chen, J. C.-Y., Saha, S., and Bansal, M. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*, 2023.

Cherian, A., Peng, K.-C., Lohit, S., Matthiesen, J., Smith, K., and Tenenbaum, J. Evaluating large vision-and-language models on children's mathematical olympiads. *Advances in Neural Information Processing Systems*, 37: 15779–15800, 2024.

Dawid, A. P. and Skene, A. M. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., and Weston, J. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*, 2023.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., and Tsvetkov, Y. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*, 2024.

Fourney, A., Bansal, G., Mozannar, H., Tan, C., Salinas, E., Niedtner, F., Proebsting, G., Bassman, G., Gerrits, J., Alber, J., et al. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

Gu, W., Han, J., Wang, H., Li, X., and Cheng, B. Explain-analyze-generate: A sequential multi-agent collaboration method for complex reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7127–7140, 2025.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., and Hovy, E. Learning whom to trust with mace. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1120–1130, 2013.

Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., and Zhou, D. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*, 2023.

Huang, M., Lai, H., Zhang, X., Wu, W., Ma, J., Zhang, L., and Liu, J. EvoChart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 2025.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, 2023.

Kamoi, R., Zhang, Y., Zhang, N., Han, J., and Zhang, R. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440, 2024.

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

Kim, G., Baldi, P., and McAleer, S. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677, 2023.

Li, J., Zhang, Q., Yu, Y., Fu, Q., and Ye, D. More agents is all you need. *arXiv preprint arXiv:2402.05120*, 2024.

Li, W., Lin, Y., Xia, M., and Jin, C. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*, 2025a.

Li, Y., Jiang, S., Hu, B., Wang, L., Zhong, W., Luo, W., Ma, L., and Zhang, M. Uni-MOE: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Shi, S., and Tu, Z. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Luo, Z., Yang, Z., Xu, Z., Yang, W., and Du, X. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.

Ma, Q. and Olshevsky, A. Adversarial crowdsourcing through robust rank-one matrix completion. *Advances in Neural Information Processing Systems*, 33:21841–21852, 2020.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Majdi, M. S. and Rodriguez, J. J. Crowd-certain: Label aggregation in crowdsourced and ensemble learning classification, 2023. URL `https://arxiv.org/abs/2310.16293`.

Pan, L., Albalak, A., Wang, X., and Wang, W. Y. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*, 2023.

Shen, L., Chen, G., Shao, R., Guan, W., and Nie, L. MoME: Mixture of multimodal experts for generalist multimodal large language models. *Advances in neural information processing systems*, 37:42048–42070, 2024.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Song, Y., Ou, T., Kong, Y., Li, Z., Neubig, G., and Yue, X. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge. *arXiv preprint arXiv:2504.10342*, 2025.

Subramaniam, V., Du, Y., Tenenbaum, J. B., Torralba, A., Li, S., and Mordatch, I. Multiagent finetuning: Self improvement with diverse reasoning chains. *arXiv preprint arXiv:2501.05707*, 2025.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

Thind, R., Sun, Y., Liang, L., and Yang, H. Optimai: Optimization from natural language using llm-powered ai agents. *arXiv preprint arXiv:2504.16918*, 2025.

Ustalov, D., Pavlichenko, N., and Tseitlin, B. Learning from crowds with crowd-kit. *arXiv preprint arXiv:2109.08584*, 2021.

Valmeekam, K., Marquez, M., and Kambhampati, S. Can large language models really improve by self-critiquing their own plans? *arXiv preprint arXiv:2310.08118*, 2023.

Wang, H., Yu, W., Du, X., Chen, Q., Chu, Z., Yan, L., Jiang, J., and Guan, Y. Multi-modal multi-agent debate-based dialectical knowledge transfer learning. *Applied Soft Computing*, pp. 113551, 2025a.

Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., and Zou, J. Mixture-of-agents enhances large language model capabilities, 2024. *URL https://arxiv. org/abs/2406.04692*.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Wang, Q., Wang, Z., Su, Y., Tong, H., and Song, Y. Rethinking the bounds of llm reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*, 2024b.

Wang, S., Liu, Z., Wei, J., Yin, X., Li, D., and Barsoum, E. Athena: Enhancing multimodal reasoning with data-efficient process reward models. *arXiv preprint arXiv:2506.09532*, 2025b.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Welleck, S., Lu, X., West, P., Brahman, F., Shen, T., Khashabi, D., and Choi, Y. Generating sequences by learning to self-correct. *arXiv preprint arXiv:2211.00053*, 2022.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Zeng, T., Zhang, S., Wu, S., Classen, C., Chae, D., Ewer, E., Lee, M., Kim, H., Kang, W., Kunde, J., et al. Versaprm: Multi-domain process reward model via synthetic reasoning data. *arXiv preprint arXiv:2502.06737*, 2025.

A. Prompts used in our MAD formulations

In Table 7, we provide the varied prompts (described in the main paper) used at various stages of our framework. Please also see the suppl.zip attachment that provides detailed solution responses, feedbacks, and orchestrator summaries for all the four datasets we used.

Prompt Type	**Prompt text**
PROMPT_1^S (Solver)	Please describe the problem, including the images if available. Your description of the image should be detailed and accurate so that an LLM without access to the image can solve the problem. After this, please solve this question with explanation of the intermediate steps. You must select an answer from one of the five options: **Puzzle:** [PROBLEM]...[/PROBLEM] **Options:** [OPTIONS] ... [/OPTIONS]. Write the final **answer option** as one of `A1, B2, C3, D4, E5`.
PROMPT^R (Reflector)	Following are the solutions of LLM in solving a math puzzle. The problem statement and the LLM solution are provided below. The problem may contain an image that needs to be used to understand the LLM's answer. You should not solve the problem yourself. Instead, please check if the solutions below are logically and mathematically consistent and the final answer is correct? [PROBLEM] ... [/PROBLEM]. [RESPONSE] ... [/RESPONSE]. Your Task: Provide a detailed explanation of your assessment so that it can be used as feedback to LLM to improve the solution, including any mistakes in understanding the given image (if any). Your response should end with a single line in the format: $FINAL_SCORE : \langle value \rangle$ where $\langle value \rangle$ is 0 if final answer is the wrong, 1 if you are unable to confirm, and 2 if the final answer is correct.
PROMPT^O (Orchestrator)	The following are the feedback received from LLMs on a puzzle solution. There could be significant differences in the feedback provided by different LLMs, especially the LLMs may select different answer options. Your task is to pay attention to these differences in the feedback and the selected answer options. Please summarize all the feedback and provide a **set of actionable questions** for the solver to revise its solution in the next round. It is possible that the previous solution might have overlooked details in the given image (if any), and thus your summary should emphasize the importance of re-evaluating the image. [FEEDBACK] [/FEEDBACK] [FEEDBACK] [/FEEDBACK] ...
$\text{PROMPT}_{k>1}^S$ (Solver)	Following is the feedback received on your previous solution. If there was a mistake in your previous solution, use this feedback and the list of actionable steps provided to reconsider and re-evaluate your solution. You should re-evaluate your understanding the given image (if any) and the problem statement using the feedback. Specifically, based on the provided feedback, please provide a detailed explanation of the given image (if any). Please provide the final answer option and explanation of the intermediate steps, as well as details of how the feedbacks were addressed in your revised final answer. Your final answer should be reported in your final line and must select one of the answer options: A1, B2, C3, D4, E5. [PROBLEM] ... [/PROBLEM] [OPTIONS] ... [/OPTIONS] [RESPONSE] ... [/RESPONSE] [FEEDBACK] ... [/FEEDBACK]

Table 7. Various prompts used in the WISE MAD formulation.

B. Ablation Studies

In Table 8, we provide publicly available information of the LLMs that we consider, their capabilities, and tasks they were trained for. Table 9 summarizes each model's accuracy, parameter size, modality capabilities, and baseline performance on the SMART-840 dataset. These scores are obtained using a single forward pass—without any debate rounds—thereby

serving as the foundational performance reference. We report results for both image–text problems and text-only problems, as our later studies incorporate text-only LLMs as reflectors that help weigh and critique the solutions proposed by multimodal LLMs.

Based on these baseline results, we group the models into strong and weak categories. The two best-performing models, GPT-4.1 and Claude-Sonnet, are designated as strong. The remaining multimodal models—GPT-4o, Gemma3, and Qwen-VL-2.5—are categorized as weak. Similarly, the two text-only models, Phi-4 and Llama3, are also considered weak for the purposes of this study. We emphasize that this classification is purely performance-driven and is intended to enable a systematic exploration of different combinations of weak and strong models as they take on varied roles in our subsequent evaluations.

LLM/LVLM	**Skills**	**Training Tasks**	Multimodal?
GPT-4o (G4o)	Reasoning; coding	Language modeling, code generation	Yes
GPT-4.1 (G41)	Coding and reasoning	Language, advanced coding tasks, reasoning	Yes
Claude-Sonnet (CS35 or S35)	Strong reasoning, coding, multilingual	General reasoning, coding, text understanding	Yes
Claude-Haiku (Claude 3)	Lightweight, efficient reasoning	General reasoning, summarization, text analysis	Yes
Qwen-VL 2.5 (QVL)	Reasoning	Multimodal reasoning, VQA, captioning, dialogue	Yes
Gemma3 (Ge3	Reasoning (presumed)	General multimodal tasks, language understanding	Yes
Phi-4 (ϕ4)	Math and coding reasoning	Language modeling, mathematics, coding	No
Llama3 (L3)	General language understanding	Language modeling, reasoning, dialogue	No

Table 8. LLMs and MLLMs that we use within the WISE architecture. We review our abbreviations and their capabilities.

B.1. Performance versus Computations:

In Table 10, we compare the performance of various model combinations within WISE under comparable compute budgets and across multiple multi-agent reasoning architectures. Specifically, we evaluate: i) Single-model aggregation, where each model independently produces one solution and the final answer is selected by majority vote; ii) Self-Reflect (Shinn et al., 2023), in which each model performs 8 rounds of self-reflection before ensembling the outputs across models using majority voting; iii) Self-consistency, similar to Self-Reflect but without feedback—each model is prompted 8 times independently and the aggregated result is used; iv) Homogeneous feedback, where each model provides feedback to its own solutions for 8 iterative rounds; and v) RECONCILE, reviewed in the main paper, a multi-agent debate framework using equally capable LLMs.

B.2. Performance on Varied LLM Configurations:

In Table 11, we evaluate a range of strong, weak, and language-only model combinations within the WISE debate framework. Specifically, we test solver–reflector pairings across five categories: strong–strong, strong–mix, mix–mix, weak–weak, and weak–language-only. Here, mix denotes a role that includes both strong and weak models. All experiments use the same orchestrator model (GPT-4o), and we evaluate performance on the SMART-840 dataset using the WISE-DS aggregation method. As expected, the strong–strong combination achieves high performance. Interestingly, the mix–mix configuration—used in our best-performing setup—achieves a slightly higher score (67.7 $\rightarrow$ 68.1), suggesting that controlled diversity across roles can be beneficial.

The weak–weak combination performs below all other pairings, as anticipated, yet it still noticeably surpasses the best individual model in the pair. For example, GPT-4o alone achieves 42.4% accuracy (Table 9), but when used in a weak–weak WISE configuration, performance rises to 49.7%—comparable to RECONCILE when using similarly weak models. Unlike

Table 9. Performance of various individual MLLMs on the SMART-840 dataset (no debate or aggregation).

LLMs / MLLMs	Size	Capability	Performance	Image+Text	Text	Overall
GPT-4.1	Unknown	VL	Strong	48.2	88.2	60.6
Claude-Sonnet	Unknown	VL	Strong	35.3	71.0	44.3
GPT-4o	Unknown	VL	Weak	33.6	67.7	42.4
Gemma3	12B	VL	Weak	24.0	73.7	38.6
Qwen-VL-2.5	7B	VL	Weak	25.0	52.5	37.2
Phi-4	3.8B	L	Weak	N/A	65.4	N/A
Llama3	8B	L	Weak	N/A	34.8	N/A

Table 10. Performance of MAD approaches on SMART-840 using varied MLLM configurations for approximately **similar number of LLM calls** on average.

Method	LLMs / MLLMs	1_2	3_4	5_6	7_8	9_10	11_12	Image+Text	Text	Overall
Single	G41+G4o+GS35+Ge3+QVL	53.3	50.0	46.7	67.3	55.3	60.7	45.5	82.7	55.6
Self-Reflect	Ge3+QVL+G41	48.3	47.5	42.0	62.0	58.7	64.7	42.6	83.4	53.9
Self-Consistency	G41+CS35+Ge3+QVL	55.0	45.8	46.0	63.3	64.0	65.3	47.2	84.1	56.6
Homogeneous	G41+CS35+Ge3+QVL	49.2	48.3	38.0	58.7	62.7	66.7	43.5	81.6	53.9
RECONCILE	G41 + Ge3 + S35	55.0	59.2	50.0	67.3	68.7	72.0	50.9	90.1	62.0
	G4o + Ge3 + QVL	50.8	41.7	36.7	54.7	56.0	62.7	39.6	76.1	50.4
	Ge3 + QVL	41.7	38.7	34.0	50.7	50.0	54.7	35.8	65.5	45.0
WISE (ours)	$(G41+Ge3+S35)^2$|G4o	**65.0**	**61.8**	**56.6**	74.3	**75.5**	**73.0**	**55.8**	**91.4**	**68.1**

Table 11. Performance of varied WISE on the SMART-840 dataset using varied configurations of the weak and strong LLM models under approximately the **same number of LLM calls.** *CH3 stands for Claude-Haiku 3.

Solvers	Reflectors	LLMs / MLLMs	1_2	3_4	5_6	7_8	9_10	11_12	Overall
Strong	Strong	$(G41+S35)^2$|G4o	66.7	59.1	58.2	78.9	71.4	72.2	67.7
Strong	Weak	$(G41+S35)\times(G4o+CH3^*+\phi4)$|G4o	59.2	58.3	49.3	73.3	60.7	69.3	61.7
Mix	Mix	$(G41+Ge3+S35)^2$|G4o	**65.0**	**61.8**	**56.6**	74.3	**75.5**	**73.0**	**68.1**
Weak	Weak	$(G4o+QVL)\times(Q+\phi4+L3+G3)$|G4o	51.4	39.4	41.0	53.7	53.9	58.7	49.7
Weak	L-only	G4o $\times$ (ϕ4+o1-mini)|G4o	49.2	45.4	47.1	54.6	56.9	55.8	51.5

RECONCILE, which assumes all participating models have similar capabilities (e.g., all text-only or all vision–language), WISE explicitly assigns models to specialized roles, enabling more flexible and heterogeneous debate setups. This flexibility becomes especially advantageous in the weak–language-only scenario, where WISE reaches 51.5% accuracy. Under similar compute budgets, this is a 1.1% improvement over RECONCILE using GPT-4o—despite operating entirely in a weak setting.

Overall, our results show that strong–strong configurations predictably benefit WISE. However, heterogeneous pairings yield mixed trends: i) strong–mix combinations where the mix role is dominated by weak reflectors tend to underperform, likely due to lower-quality feedback; ii) weak–weak and weak–L-only setups can still deliver competitive gains, suggesting that WISE 's structured debate can absorb or correct suboptimal weighting from weaker reflectors. These findings highlight the robustness and adaptability of WISE in leveraging diverse model capabilities.

B.3. Comparisons on Orchestrator Influence:

A key and novel component of WISE is the orchestrator LLM and the expanded role it plays within the multi-agent debate. In prior work (Chen et al., 2023; Bai et al., 2022; Liang et al., 2023), the orchestrator's primary function is to summarize each debate round before passing the summary to the agents in the next round. In contrast, the orchestrator in WISE performs two critical functions: i) summarizing the reflectors' feedback, and ii) generating actionable, targeted follow-up questions for the solvers. These questions help direct the solvers' attention to specific multimodal aspects highlighted by the reflectors, enabling deeper and more focused analysis in subsequent rounds. To evaluate the importance of this enhanced orchestrator role, we conducted two sets of studies: i) replacing GPT-4o (our default orchestrator) with a weaker model, Phi-4, and ii) modifying the orchestrator so that it only summarizes feedback without generating follow-up questions.

Table 12 reports the results on the SMART-840 dataset (split 1-2) using two debate configurations: strong+mix and strong–strong. Rows marked (w/o Q) correspond to the orchestrator variant that does not generate actionable questions. The results show that replacing GPT-4o with Phi-4 causes little to no performance degradation, indicating that WISE is not strongly dependent on a powerful orchestrator model. However, removing the question-generation step results in a consistent 3–4% drop in accuracy across both configurations. This clearly demonstrates that the orchestrator's ability to craft targeted follow-up questions is essential for driving effective multi-agent reasoning within WISE.

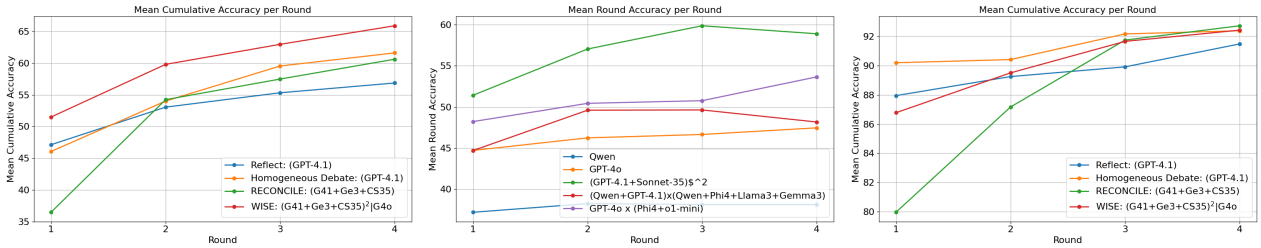
B.4. Accuracy vs Rounds:

In Figure 5, we plot the accuracy against the number of debate rounds. We plot the cumulative accuracy, which measures for a round k if the correct solution to a problem was found in some round from $1..k$. This serves as the upper-bound on the accuracy, and shows if by increasing the rounds problems can be solved. As seen in Figure 5c, the performance of the heterogeneous debate is seen to increase with the increasing number of rounds, suggesting that models that were wrong in

Table 12. Performance of varied WISE on the SMART-840 dataset with varied orchestrators configurations (using wt. majority voting).

WISE LLMs / MLLMs	Orchestrator	1_2
(G41+Ge3+S35)^2	G4o	65.0
(G41+Ge3+S35)^2	G4o (w/o Q)	62.1
(G41+Ge3+S35)^2	Phi-4	64.8
(G41+Ge3+S35)^2	Phi-4 (w/o Q)	61.2
(G41+S35)^2	G4o	66.7
(G41+S35)^2	G40 (w/o Q)	63.6
(G41+S35)^2	Phi-4	65.7
(G41+S35)^2	Phi-4 (w/o Q)	61.7

previous rounds could correct their mistakes. In Figure 5b, we plot the average accuracy over the answer selected against the ground truth. As can be seen, all the debate configurations demonstrate monotonic increase in the accuracy with the rounds. We found that using no-debate does not show much promise, while using homogeneous debate with a strong LLM (GPT-4o) closely matches the performance of the heterogeneous debate with (GPT-4o+Qwen-VL)$\times$ (Gemma3+Llama3+Phi4+Qwen-VL) with slightly higher performance than the former with more debate rounds. We also find that the stronger model (combination of GPT and Claude-3 models) perform the best, however even for them, the performance increases by nearly 10% from the first round to the third from nearly 55% to 65%.



(a) VL Performance *(b)* Acc. on Weaker Models *(c)* Text-only Performance

Figure 5. (a) plots the VL performance of various MAD setups on the SMART-840 dataset. In (b), we compare the average round performance against then number of rounds on various strong and weak LLM configurations. In (c), we plot the performance using only text-only problems. Comparing (a) and (c), we find that text-only performance of various MAD configurations are very high, and WISE closely matches the performance. However, on multimodal problems, WISE clearly wins.

B.5. Rounds vs. MAD Complexity:

In Figure 5, we provide a bar plot showing the average number of rounds taken by a debate configuration to reach consensus. Recall that we stop the debate when the orchestrator finds that all the Solvers produce the same answer and all the Reflectors agree that the solution is correct, even when the predicted solution may not match the ground truth. As can be seen from the plot, having more LLMs in the debate (e.g., the red bar) demonstrates a significantly longer number of rounds to converge (nearly 4 rounds) while the rounds average around 1.5 for the strong configuration (green bar). We show round = 1 for single attempt and repeated sampling variants. We also find that the number of debate rounds is lower for the homogeneous variant, suggesting the lack of ability in the model to understand its own mistakes to be corrected (homogeneous configuration provides feedback on its own previous solution).

B.6. Performance on Text-Only Problems:

In Figures 5c, we plot the summarized performance for all text-only problems in the SMART-840 dataset. From the figure, we find the performance of the models on text-only problems is nearly 20% higher than on image-text + text-only problems – suggesting the strong models are almost close to solving all the text-only problems in the dataset correctly (which amounts to nearly 30% of the problems). Despite this significant performance, we find that homogeneous models (that includes GPT-4o) still lags below by nearly 16% (73% -> 89%) when combined with Claude-3, and that the gap between the cumulative and the average performance on this subset if merely 4%, suggesting the debate is able to near closing the gap, yet the WISE debate is useful.

C. WISE-DS Expectation Maximization

Repeating the approach for WISE-Dawid-Skene given in the main paper, we provide more details here on how the ensued Expectation Maximization formulation is derived and how it is solved.

C.1. Derivation of EM for WISE-DS:

As noted, given N worker responses, each selecting one of K possible answers for a given problem, the classic Dawid–Skene (DS) algorithm employs expectation–maximization (EM) to jointly estimate the error rates of workers and the latent ground-truth label distribution. In our context, assuming conditional independence of agents' responses, we cast our MAD formulation within the DS framework to model agent-specific error rates and compute the posterior over solution responses. This allows us to down-weight agents that behave randomly (or bluffing) while amplifying the influence of those whose outputs are consistently aligned with the majority.

Suppose $p_{\alpha\beta}$ is the probability of selecting answer β for the true answer α by a model, and if ζ_α is the true prior probability of selecting an option α, then given Λ problems and the agents' responses, (Dawid & Skene, 1979) models the likelihood of the data as a mixture of multinomial distributions:

$$p(\text{data}) \propto \prod_{i=1}^{\Lambda} \sum_{\alpha=1}^{K} \zeta_\alpha \left[\prod_{j=1}^{N} \prod_{\beta=1}^{K} (p^j_{\alpha\beta})^{\lambda^j_{i\beta}} \right], \tag{5}$$

where $\lambda^j_{i\beta}$ is the number of times j-th agent produced β as the answer against the true answer of α if the model is run multiple times on the same problem i. As both the true priors ζ_γ and the error rates $p_{\alpha\beta}$ are unknown, DS uses EM to optimize the likelihood iteratively towards convergence.

In WISE, we have two sources of errors: i) solvers making errors in selecting the correct answers from the K choices and the reflectors selecting one of $J(=3)$ weights. Suppose $p^t_{\alpha_1\beta_1}$ and $p^c_{\alpha_2\beta_2}$ denote the error matrices for the solver and the reflectors respectively, in selecting option $\beta_1 \to \alpha_1$ (for $\alpha_1, \beta_1 \in [K]$) and selecting wrong weights for a provided answer, i.e., selecting weight $\beta_2 \to \alpha_2$ (for $\alpha_2, \beta_2 \in [J]$), then the joint log-likelihood for estimating the combined error matrices and their joint true priors $\zeta_{\alpha\beta} = \zeta_\alpha \zeta_\beta$ could be modeled for debate round k as the product of two mixtures of multinomial distributions, given by:

$$\log \mathcal{L} = \sum_{i=1}^{\Lambda} \log \sum_{\alpha=1}^{J} \sum_{\beta=1}^{K} \zeta_\alpha \zeta_\beta \prod_{t,\beta_1} \left(P^t_{\beta,\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \prod_{c,\beta_2} \left(P^c_{\alpha,\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}} \tag{6}$$

$$= \sum_{i=1}^{\Lambda} \log \sum_{\alpha=1}^{J} \sum_{\beta=1}^{K} \zeta_\alpha \zeta_\beta \prod_{t,\beta_1} \left(P^t_{\beta,\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \prod_{c,\beta_2} \left(P^c_{\alpha,\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}}. \tag{7}$$

where $\alpha \in \{1, \ldots, J\}$ denote the latent true *rank/weight* label and $\beta \in \{1, \ldots, K\}$ denote the latent true *solution/answer* label. We assume factorized priors

$$\zeta_{\alpha\beta} = \zeta_\alpha \, \zeta_\beta, \qquad \text{s.t.} \sum_{\alpha=1}^{J} \zeta_\alpha = \sum_{\beta=1}^{K} \zeta_\beta = 1. \tag{8}$$

Each *solver* $t \in \mathcal{S}$ emits an answer $\beta_1 \in \{1, \ldots, K\}$ with confusion matrix $P^t_{\beta,\beta_1} = P(\text{emit } \beta_1 \mid \text{true } \beta)$, and each *critic/reflector* $c \in \mathcal{R}$ emits a rank $\beta_2 \in \{1, \ldots, J\}$ with confusion matrix $P^c_{\alpha,\beta_2} = P(\text{emit } \beta_2 \mid \text{true } \alpha)$. For item i, let $\lambda^{S_t}_{i\beta_1}$ and $\lambda^{R_c}_{i\beta_2}$ denote the observed counts of solver and critic emissions, respectively.

Given independence of solvers and critics conditioned on (α, β), the likelihood of observations for item i is

$$P(\text{obs}_i \mid \alpha, \beta) = \prod_{t \in \mathcal{S}} \prod_{\beta_1=1}^{K} \left(P^t_{\beta,\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \cdot \prod_{c \in \mathcal{R}} \prod_{\beta_2=1}^{J} \left(P^c_{\alpha,\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}}. \tag{9}$$

The observed-data likelihood over Λ items is therefore

$$\mathcal{L}(\Theta) = \prod_{i=1}^{\Lambda} \sum_{\alpha=1}^{J} \sum_{\beta=1}^{K} \zeta_\alpha \zeta_\beta \; P(\text{obs}_i \mid \alpha, \beta), \tag{10}$$

where $\Theta = \{\zeta_\alpha, \zeta_\beta, P^t, P^c\}$.

E–step. For each item i, the posterior responsibility of the latent pair (α, β) is given by

$$\gamma_{i,\alpha\beta} = P(\alpha, \beta \mid \text{obs}_i) = \frac{\zeta_\alpha \, \zeta_\beta \left[\prod_{t \in \mathcal{S}} \prod_{\beta_1=1}^{K} \left(P^t_{\beta,\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \right] \left[\prod_{c \in \mathcal{R}} \prod_{\beta_2=1}^{J} \left(P^c_{\alpha,\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}} \right]}{\sum_{\alpha'=1}^{J} \sum_{\beta'=1}^{K} \zeta_{\alpha'} \, \zeta_{\beta'} \left[\prod_{t \in \mathcal{S}} \prod_{\beta_1=1}^{K} \left(P^t_{\beta',\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \right] \left[\prod_{c \in \mathcal{R}} \prod_{\beta_2=1}^{J} \left(P^c_{\alpha',\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}} \right]}.$$

The corresponding marginal posteriors for the true solution and rank are obtained by summing over the other latent variable:

$$p^t_i(\beta) = \sum_{\alpha=1}^{J} \gamma_{i,\alpha\beta}, \qquad p^c_i(\alpha) = \sum_{\beta=1}^{K} \gamma_{i,\alpha\beta}. \tag{11}$$

M–step. Define the expected counts

$$N_{\alpha\beta} = \sum_{i=1}^{\Lambda} \gamma_{i,\alpha\beta}, \quad N_\alpha = \sum_\beta N_{\alpha\beta}, \quad N_\beta = \sum_\alpha N_{\alpha\beta}. \tag{12}$$

The factorized priors are updated as

$$\zeta^{\text{new}}_\alpha = \frac{N_\alpha}{\sum_{\alpha'} N_{\alpha'}}, \qquad \zeta^{\text{new}}_\beta = \frac{N_\beta}{\sum_{\beta'} N_{\beta'}}. \tag{13}$$

The solver and critic confusion matrices are updated as weighted multinomial maximum likelihoods:

$$P^{t,\text{new}}_{\beta,\beta_1} = \frac{\sum_i \left(\sum_\alpha \gamma_{i,\alpha\beta}\right) \lambda^{S_t}_{i\beta_1}}{\sum_{\beta'_1} \sum_i \left(\sum_\alpha \gamma_{i,\alpha\beta}\right) \lambda^{S_t}_{i\beta'_1}}, \tag{14}$$

$$P^{c,\text{new}}_{\alpha,\beta_2} = \frac{\sum_i \left(\sum_\beta \gamma_{i,\alpha\beta}\right) \lambda^{R_c}_{i\beta_2}}{\sum_{\beta'_2} \sum_i \left(\sum_\beta \gamma_{i,\alpha\beta}\right) \lambda^{R_c}_{i\beta'_2}}. \tag{15}$$

Posteriors at convergence. Let $\widehat{\zeta}_\alpha$, $\widehat{\zeta}_\beta$, $\widehat{P}^t$, and $\widehat{P}^c$ denote the parameters at EM convergence. For item i, the joint posterior over the latent pair (α, β) is

$$\widehat{\gamma}_{i,\alpha\beta} = P(\alpha, \beta \mid \text{obs}_i; \widehat{\Theta}) = \tag{16}$$

$$\frac{\widehat{\zeta}_\alpha \, \widehat{\zeta}_\beta \left[\prod_{t \in \mathcal{S}} \prod_{\beta_1=1}^{K} \left(\widehat{P}^t_{\beta,\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \right] \left[\prod_{c \in \mathcal{R}} \prod_{\beta_2=1}^{J} \left(\widehat{P}^c_{\alpha,\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}} \right]}{\sum_{\alpha'=1}^{J} \sum_{\beta'=1}^{K} \widehat{\zeta}_{\alpha'} \, \widehat{\zeta}_{\beta'} \left[\prod_{t \in \mathcal{S}} \prod_{\beta_1=1}^{K} \left(\widehat{P}^t_{\beta',\beta_1}\right)^{\lambda^{S_t}_{i\beta_1}} \right] \left[\prod_{c \in \mathcal{R}} \prod_{\beta_2=1}^{J} \left(\widehat{P}^c_{\alpha',\beta_2}\right)^{\lambda^{R_c}_{i\beta_2}} \right]}. \tag{17}$$

The marginal posteriors for the true solution (solver space) and the true rank/weight (critic space) are then

$$\widehat{p}_i^t(\beta) = \sum_{\alpha=1}^{J} \widehat{\gamma}_{i,\alpha\beta}, \qquad \widehat{p}_i^c(\alpha) = \sum_{\beta=1}^{K} \widehat{\gamma}_{i,\alpha\beta}. \tag{18}$$

The corresponding MAP estimates are

$$\widehat{\beta}_i^{\text{MAP}} = \arg \max_{\beta \in \{1,\ldots,K\}} \widehat{p}_i^t(\beta), \tag{19}$$

$$\widehat{\alpha}_i^{\text{MAP}} = \arg \max_{\alpha \in \{1,\ldots,J\}} \widehat{p}_i^c(\alpha). \tag{20}$$